A 32-Band Sub-band/Transform Coder Incorporating Vector Quantization for Dynamic Bit Allocation

C. D. Heron
R. E. Crochiere
R. V. Cox

Acoustic Research Department
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

In this paper we report on a study of a technique for 32-band sub-band/transform coding at 16 kb/s. This approach occupies the middle range of algorithm complexities and frequency resolution between that of Sub-Band Coding (SBC) and Adaptive Transform Coding (ATC). Two designs for 16 kb/s 32-band coders have been simulated on a laboratory computer. The results of informal listening tests indicate that the new designs offer performance comparable to existing ATC techniques while having complexities roughly three times that of existing 4 and 5 band sub-band coders.

I. Introduction

Recently frequency domain techniques for coding digital voice have received considerable attention. An area not widely explored has been the middle range of techniques such as systems with 16 to 64 frequency bands which have complexities ranging between those of sub-band and transform coders [1,5]. We have sought to answer the question of how effectively these systems can perform and what complexities are involved for coding digitized voice at bit rates of 16 kb/s.

The 32-band coders discussed in this paper incorporate a hybrid of ideas from both SBC and ATC techniques. The technique accomplishes a 32-band decomposition of speech in which bit assignment patterns and step-sizes for encoding and decoding each sub-band are dynamically updated on the basis of a new spectral side-information model. This model is based on recent concepts of template based bit assignment selection [17] and vector quantizing [6-8] in conjunction with known clustering methods previously applied to speech recognition [9,10] and vocoding.

II. The 32-Band Sub-band/Transform Coder Algorithm

Figure 1 is a block diagram for the basic 32-band sub-band/transform coder design we have studied. The division of the voice signal into sub-bands is achieved by a quadrature mirror filter (QMF) bank [11,12] discussed shortly. Each sub-band signal (shown in the figure as $y_1(m)$, through $y_{32}(m)$) is independently coded and decoded using an APCM scheme incorporating a dynamically varying bit assignment and a separate quantizer step-size in each band. The bit assignment and quantizer step-size vectors (show in the figure as $\bar{b}(m)$ and $\bar{\Delta}(m)$ respectively) for sample time m are provided by a spectral side-information model of the input speech. The model represents the time-varying spectral envelope of the input speech by a switched source of LPC spectral templates in accordance with recent concepts of vector quantization. We have experimented with a set of spectral templates ranging from 4 to 32 in number and from 2 to 10 in LPC order. The ith template is represented by the LPC vector $1/A_i(z)$. Each template that is selected to model a segment of input speech defines a one-to-one mapping to a bit assignment pattern in the bit assignment codebooks for encoding and decoding that speech segment. The best LPC template choice, i, is updated at 16 msec intervals and results in the minimum energy residual of the signals obtained from filtering the input speech segment with the ith inverse LPC filter $A_i(z)$. This method can be shown to be equivalent to minimizing the log likelihood ratio distance measure between the template spectrum and the best gain-normalized LPC model (of the same order) of the speech spectrum. The square root of the minimum residual energy, $\sqrt{\sigma_i}$, is the gain term of the best LPC template (denoted in the figure by the codeword $c(m)$) and is used with that template for establishing the quantizer step-sizes for encoding and decoding each band.

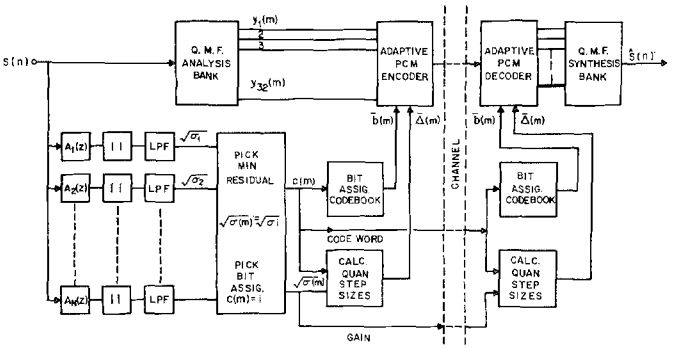

Fig. 1 32-band sub-band/transform algorithm

III. The Filter Bank

Sub-band decomposition based on a tree arrangement of QMF's has been shown to avoid the aliasing effects due to finite filter transition bandwidths and decimation [11]. The analysis and synthesis filter banks of Fig. 1 were derived from a 5-level uniform tree structure design of QMF's. The actual implementation of the filter banks was achieved with an equivalent parallel structure of band-pass filters according to the technique proposed in [12]. The quadrature mirror filters used in this structure were designed by J. D. Johnston [13] and were 64, 32, 24, 16, and 12 tap filters respectively in stages 1 through 5.

IV. A Spectral Side-Information Model Based on a Switched Source of All-pole Spectral Templates

LPC spectral templates are derived through an iterative clustering algorithm which involves spectral comparisons on a large set of LPC training vectors obtained by autocorrelation analysis on a speech data base. The clustering algorithm employed was developed by B. H. Juang [8] and implemented by L. R. Rabiner. The procedure identifies a set of LPC template vectors as cluster centers such that an average spectral distortion measure — in this case the log likelihood ratio — from all input vectors to their best match in the template vector collection is minimized.

The training database was comprised of speech by 7 speakers, 4 male and 3 female. Each speaker contributed about 19.6 seconds of speech by uttering the same 10 phonetically balanced sentences. This resulted in a 137 second database. The majority of silence was eliminated from this database to ensure that numerous shades of background noise did not occupy valuable space as spectral models. The speech used was high quality microphone speech bandpass filtered from 100 to 3500 Hz and sampled at 8 kHz.

Since the number of LPC template vectors used to represent the input speech is very small, the spectral distances between any two of these vectors will necessarily be considerable in order to encompass a variety of speech sounds. This has been an important consideration for developing an efficient method for selecting the minimum energy residual necessary for template codeword and gain selection. The lower left hand corner of Fig. 1 depicts such a method in which the square root of the nth residual energy, $\sqrt{\sigma_n}$, is estimated by low-pass filtering the absolute value of the nth residual signal. This signal is

CH1841-6/83/0000-1276 $1.00 © 1983 IEEE

obtained from filtering a speech segment with the nth inverse template filter $A_n(z)$. These computations are executed for each of the N inverse template filters for each segment of speech in order to select that template i with the minimum gain estimate $\sqrt{\sigma_i}$. Since speech is approximately stationary over 16 ms intervals, the bit assignment pattern and quantizer step-sizes for coding the sub-bands need only be updated after every fourth 4-ms frame; hence the side-information (consisting of the gain and index value) need only be transmitted every 4 frames (i.e., every four values of m).

V. Bit Assignment

The bit assignment scheme determines the number of bits that are allocated for quantizing each sub-band signal, and hence for the off-line calculation of the bit assignment codebooks of Fig. 1. Bit assignment is based on a local estimate of the amplitude variance of the signal in each sub-band which is provided from the spectral template used to represent the speech. The kth sub-band is uniformly quantized with step-size Δ_k and 2^{b_k} quantizer levels (obtained from the kth terms of the vectors $\bar{\Delta}(m)$ and $\bar{b}(m)$).

If the sub-band output sequence $y_k(m)$ of the kth sub-band is modeled by a sequence of stationary Gaussian random variables with variance γ_k^2 (the estimate of γ_k^2 is provided by the estimate of the spectral magnitude, obtained from template i, in band k), and if an average mean-squared quantization error D^* is minimized, the optimal bit assignment for the kth sub-band signal is given by [14]

$$b_k = \delta + \frac{1}{2}\log_2 \frac{\gamma_k^2}{D^*} . \tag{1}$$

The term δ is a correction term that accounts for the performance of practical quantizers. D^* is the average mean-squared distortion or quantization noise variance given by

$$D^* = \frac{1}{32}\sum_{k=1}^{32} \gamma_{e,k}^2 \tag{2}$$

where $\gamma_{e,k}^2$ is the noise variance which results from quantizing the kth output $y_k(m)$. The parameter D^* must be chosen such that

$$B = \sum_{k=1}^{32} b_k . \tag{3}$$

That is, the sum of the bits available for quantizing each channel is equal to the total number of bits B available for quantizing each frame. At a 16 kb/s coding rate there are 250 frames/second and hence B is roughly 64 bits (neglecting requirements for transmission of side-information of gain and template number). It has been shown [14] that the above bit assignment rule leads to a flat noise distribution in frequency, i.e., $\gamma_{e,k}^2 = D^*$ for all bands.

VI. Issues of Adaptive Quantization of the Sub-band Signals

The quantizer step-size Δ_k in band k, (the kth element of $\bar{\Delta}(m)$) has been determined experimentally according to the following relationship

$$\Delta_k = \frac{C \cdot \sqrt{\sigma_i}}{(2^{b_k-1})|A_i(\omega_k)|} \tag{4}$$

where $1/|A_i(\omega_k)|$ is the gain normalized spectral magnitude of the ith optimum LPC template in the kth band, $\sqrt{\sigma_i}$ is the gain (minimum residual energy) term, b_k is the number of bits allocated for quantizing the kth band, and C is an experimentally determined parameter (the quantizer loading factor).

In the first attempts at coding, eight fourth-order LPC templates were used as spectral models and C was chosen to be 2.0. The sub-band signal $y_k(m)$ was quantized with a uniform quantizer with a mid-rise characteristic when at least 1 bit was allocated to the quantizer. Otherwise the quantizer output was forced to zero. The major problem with this technique is that bands receiving only one or two bits have step-sizes which are often too large. The dominant reason for this is that the LPC templates generally over-estimate the magnitude of the spectrum in low amplitude regions. Because the number of bits assigned to a band is proportional to the amplitude of the template model in this band, bands receiving a small number of bits correspond to low amplitude bands in the original speech. However, since we are trying to model the spectral magnitude of a signal which exhibits considerable dynamic range with a small set of spectral models, and since all-pole models are used, the spectral peaks are modeled more accurately than the low amplitude regions. The problem is further exacerbated because the dynamic range of a quantizer is proportional to the number of bits it receives. Hence quantizers receiving one or two bits do not have the dynamic range to capture the overestimated signal. Fig. 2 is a magnitude plot of the output range of the quantizers in each band (the dotted lines) superimposed over the signal level in each band (the solid lines).

This problem is manifested by the presence of a trailing "noise shadow" in the coded output. That is, when the input to the quantizers in certain frequency bands lies outside the dynamic range of the quantizers a large amount of spectral error results in some channels. This spectral error manifests itself in the time domain as a trailing noise due to the long (514-sample) impulse responses of the synthesis filter bank. This poor time resolution, caused by the time-frequency duality principle (i.e., by high frequency resolution), causes gross quantizing errors in frequency to be dragged out in time for the duration of the impulse responses of the QMF filters.

In order to achieve major improvements in the coded speech quality the complexity of the spectral side-information model must be increased. Considerable improvement has been realized by the use of 32 10th-order LPC templates. An increased number of templates allows less generality of the models, while an increased LPC order offers better resolution of spectral peaks and valleys. Fig. 3 illustrates the increased accuracy of coding with 32 10th-order templates. The figure includes a standard 14th-order LPC analysis of the input and coded speech, the template used, and the bit assignment.

There is a practical limit to the amount of improvement which can be bought with additional templates and LPC order. Listening tests performed with 64 and 128 (10th-order LPC) templates yielded only marginal improvement in subjective quality over using 32, 10th-order LPC templates. Although more templates result in some improvement in the bit assignment patterns, there are never enough templates to do good modeling in the 1 and 2 bit bands which correspond to low amplitude regions in the input spectra.

A substantial improvement in the quality of the coded speech results from reducing the quantizer loading factor when quantizing with only one or two bits. This has the effect of proportionally decreasing the step-sizes of these quantizers so they can better capture the input signal. The best results have been achieved by lowering the step-sizes of 1 and 2 bit quantizers by approximately 9 and 5 dB respectively. This is implemented by changing C in Eq. (4) to 0.665 and 1.054 for 1 and 2 bit quantizers respectively.

By minimizing the quantizer error variance for the case of an equally spaced level quantizer with a normally distributed input signal, Max [15] found that the degree of quantizer loading should increase as the number of quantizer levels increase. As a final improvement to the step-size adaptation scheme, quantizers receiving 3,4, or 5 bits used optimum step-sizes according to Max. This was accomplished with C parameters of 2.344, 2.682, and 3.01 for 3,4, and 5 bit quantizers respectively. This technique offers an increased amount of dynamic range in large amplitude portions of the spectra.

Fig. 4 shows a magnitude plot of the quantizer output range (the dotted lines) superimposed over the signal level (the solid lines) for each band using the final step-size adaptation scheme. Comparison of Fig. 4 with Fig. 2 shows the improvement in capturing the input signal resulting from:

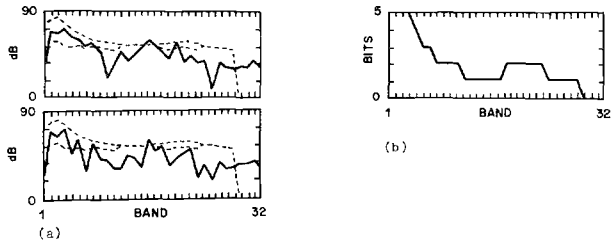

Fig. 2 a) Successive 4 ms frames of the QMF analysis filter bank output (solid line) superimposed over minimum and maximum quantizer levels,

b) the corresponding bit assignment.

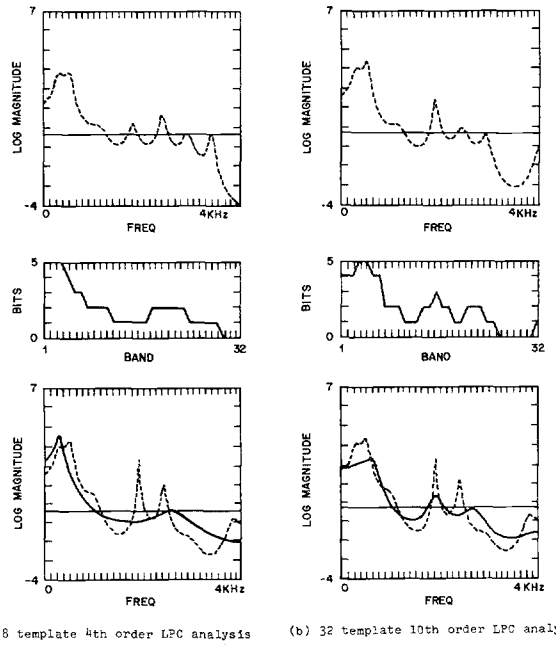

Fig. 3 Illustrations (from top to bottom) of: the LPC spectrum of the coded output, the bit assignment, and the LPC input spectrum (dotted line) superimposed with the template LPC spectrum.

1. More templates of higher LPC order
2. Subjectively reducing the step-sizes for quantizers receiving 1 or 2 bits
3. Increasing the step-size loading according to Max for quantizers receiving 3,4 or 5 bits.

In this scheme side-information is transmitted once for every four 4-ms frames without any noticeable degradation in coded speech quality. This scheme corresponds to 12 bits (five for the template number and 7 for the gain) transmitted every four frames leaving 61 bits for quantizing each 32-sample frame. This side-information model requires 750 b/s from the total 16 kb/s coding rate.

A second quantizing scheme, incorporating dynamic pre-emphasis of the signal prior to coding, has been attempted in order to reduce the effects of energy leakage from large energy bands to small energy bands. An adaptive pre-emphasis filter is used to obtain a spectrally flattened output prior to analysis/synthesis. The pre-emphasis filter is the inverse LPC template filter $A_i(z)$ of Fig. 1 which best models the input speech in the current frame. The output of the QMF synthesis filter bank is filtered with the corresponding LPC synthesis template filter $1/A_i(z)$ in order to de-emphasize the coded signal. The spectral side-information model consists of the same 32 10th-order LPC template selection procedure as in the previous scheme and requires 750 b/s from the total 16 kb/s coding rate.

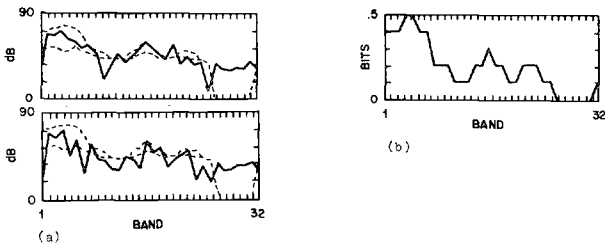

Fig. 4 a) Successive 4 ms frames of the QMF analysis filter bank output (solid line), superimposed with minimum and maximum quantizer output levels.

b) The corresponding bit assignment.

VII. Results of Informal A versus B Comparison Testing

An informal comparison test for quality was performed to compare the effectiveness of both 32-band sub-band coder designs with adaptive transform coding, and several previous sub-band coding techniques, all at bit-rates of 16 kb/s. Seven different frequency domain coding techniques were included: 1) the 32-band sub-band algorithm without dynamic pre-emphasis, 2) the 32-band sub-band algorithm with dynamic pre-emphasis, 3) a non-uniform four-band SBC algorithm [5] 4) an ATC algorithm using a homomorphic side-information model [3], 5) a four-band integer-band sampling SBC algorithm [4] , 6) a five-band octave-band SBC algorithm [16], 7) a five-band octave band SBC algorithm incorporating pitch prediction [16]. The numbers used to denote the various methods are listed at the bottom of Table 1.

Each coder was compared against all other coders and the original source material in a forced choice *A* versus *B* decision for quality. The source material consisted of eight different phonetically balanced sentences. Two male and two female speakers were used in the test. A randomized mix of sentence material was used and not all of the sentences were used to compare each coder to every other coder. However, each coder was compared against every other coder for at least two different sentences such that an *A* versus *B* comparison was accompanied by a corresponding *B* versus *A* comparison (in randomized order) to reduce any listener bias towards the first or second sentence of an *A-B* pair. A total of 24 listeners were used in the experiment. Two different randomized orders were used with 12 listeners each. Each listener made a total of 56 comparisons and each coder was compared against every other coder twice for each listener yielding a total of 48 comparisons with every other coder.

Table 1 shows a ranking of the coding algorithms according to the percentage of total possible votes that they could receive in the experiment (the total number of comparisons made in the experiment for each coder against all other coders or originals was 336). As seen in the table the original was distinctly preferred in terms of quality in 94 percent of its comparisons. This was followed by the ATC algorithm of [3] and the 32-band sub-band algorithms without and with dynamic pre-emphasis respectively.

A second experiment was performed in order to obtain more detailed information about the relative performance of those algorithms most preferred. Four coding techniques were included in this test corresponding to coder numbers 1,2,3, and 4. The conditions of this experiment were identical to those of the first experiment with the following exception. Each listener made a total of 40 comparisons and each coder was compared against every other coder four times for each listener yielding a total of 96 comparisons with every other coder.

Table 2 shows percentage listener preferences for the different *A* versus *B* comparisons. For example the lower right entry indicates that in comparing coder #4 to coder #3, 78 percent of the listeners chose coder #4 and 22 percent chose coder #3.

Table 3 shows a ranking of the coding algorithms according to the percentage of total possible votes that they could receive in the experiment (the total number of comparisons made in the experiment for each coder against all other coders on originals was 384).

TABLE 1

Ranking of Coders According to Percent of *A* versus *B* Votes Received

Coder	Percent of Total Possible Coder Votes Received
0	94
4	71
1	60
2	54
7	44
3	42
6	23
5	13

CODER NUMBERS

#0 - Original (uncoded)

#1 - 32-band Sub-band algorithm without dynamic pre-emphasis at 16 kb/s

#2 - 32-band Sub-band algorithm with dynamic pre-emphasis at 16 kb/s

#3 - Non-uniform 4 band SBC [5] at 16 kb/s

#4 - ATC algorithm using the homomorphic side information model [3] at 16 kb/s

#5 - Four-band integer-band sampling SBC algorithm [4] at 16 kb/s

#6 - Five-band octave band SBC algorithm [16] at 16 kb/s

#7 - Five-band octave band SBC algorithm incorporative pitch prediction [16] at 16 kb/s.

TABLE 2 Percent Preference of Coders In an *A* versus *B* Comparison

		B			
	A vs B	#0	#1	#2	#3
A	#1	1:99			
	#2	5:95	36:64		
	#3	3:97	19:81	38:62	
	#4	6:94	52:48	70:30	78:22

TABLE 3 Ranking Of Coders According to Percent of *A* versus *B* Votes Received

Coder	Percent of Total Possible Votes Received
0	96
4	52
1	48
2	35
3	20

As shown by the overall preference ratings of Table 3, the 32-band sub-band scheme without dynamic pre-emphasis is comparable in quality to the ATC algorithm of [3] at the same bit rate. In addition both 32-band sub-band coders (with and without dynamic pre-emphasis) were judged as having better quality than the four-band SBC of [5].

Although the 32-band sub-band coder without dynamic pre-emphasis scored 13 percent better in overall preference ratings than the 32-band sub-band coder with pre-emphasis, both schemes sounded similar. Due to the subtle differences in the distortion characteristics of the two coders, people generally preferred the scheme without dynamic pre-emphasis. Nevertheless, the fact that dynamic pre-emphasis did not improve the subjective performance of the 32-band sub-band design indicates that this scheme is not worth the additional processing required. This strongly suggests that the amount of internal frequency domain aliasing of the filter bank does not have a significant effect for the filter bank used.

VIII. Conclusions

In this paper we have explored techniques of frequency domain coding of digital voice with complexities and frequency resolutions which occupy the middle range between traditional sub-band and adaptive transform coding methods. Attention has focused on studying the performance of such systems at bit rates of 16 kb/s. Two new techniques for 32-band sub-band/transform coding at 16 kb/s have been introduced. Dynamic bit assignment and quantizer step-size adaptation have been accomplished with a spectral side-information model consisting of a switched source of all-pole spectral templates in accordance with recent concepts of vector quantizing.

The basic performance limitation of the new coders results from the limited accuracy of the spectral side-information model as well as the poor time resolution of the filter bank used for reconstructing the sub-band signals.

Preliminary results of simulations and informal subjective testing indicate that one of the 32-band sub-band/transform coder designs offers comparable performance with the adaptive transform coder. This was achieved at roughly three times the algorithm complexity of the 4-band SBC design.

REFERENCES

[1] R. E. Crochiere, S. A. Webber, J. L. Flanagan, "Digital Coding of Speech in Sub-Bands," BSTJ, pp. 1069-1085, October 1976.

[2] J. M. Tribolet, R. E. Crochiere, "Frequency Domain Coding of Speech," IEEE Trans. ASSP, pp. 512-530, October 1979.

[3] R. V. Cox, R. E. Crochiere, "Real-time Simulation Of Adaptive Transform Coding," IEEE Trans. ASSP, pp. 147-154, April 1981.

[4] R. E. Crochiere, "On the Design of Sub-band Coders for Low-Bit-Rate Speech Communication," BSTJ, pp. 747-770, May-June 1977.

[5] R. E. Crochiere, R. V. Cox, J. D. Johnston, "Real Time Speech Coding," IEEE Trans. Commun., April 1982.

[6] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech Coding Based Upon Vector Quantization," IEEE Trans. ASSP, pp. 562-574, October 1980.

[7] D. Y. Wong, B. H. Juang and A. H. Gray, Jr., "Recent Development in Vector Quantization for Speech Processing," Proc. IEEE Int. Conf. ASSP (March 1981), pp. 1-4.

[8] B. H. Juang, D. Y. Wong, A. H. Gray, Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding," IEEE Trans. ASSP, pp. 294-304, April 1982.

[9] L. R. Rabiner, and J. G. Wilpon, "Considerations in Applying Clustering Techniques to Speaker-Independent Word Recognition," J. Acoust., Soc. Am., pp. 663-673, September 1979.

[10] S. E. Levinson, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition," IEEE Trans. ASSP, pp. 134-141, April 1979.

[11] D. Esteban and C. Galand, "Application of quadrature mirror filters to split band voice coding schemes," ICASSP, Hartford, CT, pp. 191-195, May 1977.

[12] D. Esteban and C. Galand, "Parallel approach to quasi perfect decomposition of speech in sub-bands," Ninth International Congress on Acoustics, Madrid, April 1977.

[13] J. D. Johnston, "A Filter Family Designed for Use in Quadrature Mirror Filter Banks," Proc. IEEE Int. Conf. ASSP (April 1980), pp. 219-4.

[14] J. Huang and P. Schultheiss, "Block Quantization of Correlated Gaussian Random Variables," IEEE Trans. Commun. Syst., Vol. CS-11, pp. 289-296, 1963.

[15] J. Max, "Quantizing for Minimum Distortion," IRE Trans. Inform. Theory, Vol. IT-6, pp. 7-12, March 1960.

[16] A. J. Barabell, *An Analysis of Sub-Band Coding Techniques For Speech Communication*, MSEE Thesis, M.I.T., Cambridge, MA, October 1981.

[17] J. D. Tomcik, Private Communication